Preliminary Reflections on a Moral Turing

Gerdes, Anne; Øhrstrøm, Peter

Published in:
ETHICOMP 2013 Conference Proceedings

Publication date:
2013

Document Version
Early version, also known as pre-print

PRELIMINARY REFLECTIONS ON A MORAL TURING TEST

Anne Gerdes and Peter Øhrstrøm

Abstract
In the wake of the growing interest in human robot interaction, it might be fruitful to explore artificial moral agency by reflecting upon the possibility of a Moral Turing Test (MTT); and whether its obvious lack of focus on interiority, i.e., its behaviouristic foundation, counts as an obstacle to establishing a test to judge the performance of an Artificial Moral Agent (AMA). Subsequently, in order to settle whether a MTT could serve as a useful framework for the understanding, design and engineering of AMA's, we set out by addressing fundamental challenges within the field of robot ethics regarding the formal representation of moral theories and standards. Here typically three design approaches to Artificial Moral Agents are available; top down theory-driven models and bottom-up approaches which set out to model moral behaviour by means of models for adaptive learning, such as neural networks. And finally, hybrid models, which involve components from both top-down and bottom-up approaches to the modelling of moral agency. With inspiration from Allen and Wallace (2009, 2000) as well as A.N. Prior (1949, 2003), we elaborate on theoretical driven approaches to machine ethics by introducing deontic tense logic. Finally, within this framework, we reflect upon the character of human interaction with a robot, which has successfully passed a MTT.

Keywords
Moral Turing Test, Artificial Moral Agents (AMAs), robot ethics, human robot interaction, inner states, deontic tense logic, branching time

1. "As if" – A Moral Turing Test

"Since being challenged to further activity, being set greater obstacles to overcome, is the sum and substance of our lives as teleological beings, developing robots – setting ourselves further technological-cultural goals – is not an inhuman or antihuman enterprise. It is simply part and parcel of the life of a species that first began cultivation the land, devising tools and machines, and cultivation – culturally developing – members of the species itself. Machines and artefacts are an inevitable part of human culture. Moral robots are merely a part that still lies in the future." (Versenyi, 1974, p. 259)

Due to the growing interest in human robot interaction (e.g. Benford & Malartre, 2007; Dautenhahn, 2007; Turkle, 2011; Wilkes, 2010; Levy, 2008 and the Geminoid Lab of Henrik Schärfe); it might be fruitful to discuss artificial moral agency by considering the possibility of a Moral Turing Test (MTT), which might enable us to distinguish principles for evaluating morally correct *actions* rather than (as in the originally Turing test (1950)) skills of articulation. The Turing test is based on a criterion of indistinguishability, meaning that a computer system passes the test if a human interrogator is unable to distinguish between utterances produced by the computer and those produced by a human. It is important to point out that the development of a system that can pass the MTT will only be a first step towards producing an AMA. The kind of machine ethical reasoning needed in order to pass the MTT should not be confused with ethical autonomous decision making. According to McDermott (2008), ethical decision making involves a conflict between self-interest and ethics, whereas challenges regarding ethical reasoning concern how to formalize human reasoning processes, which are based on moral principles and may be computationally very complex, although they are not structurally different from other kinds of reasoning processes (McDermott, 2008, p. 2). Ethical reasoning presupposes a notion of free choice in the sense that alternative future possibilities are open to the persons in question. In order to be a genuine ethical decision maker one must also be free in the sense that one can sometimes choose to act in one's self-interest even though it runs counter to moral prescriptions. A robot does not have to be free in this sense in order to pass a MTT.

Given that we depend on various kinds of robot services, then at least for the sake of safety, we may want AMAs to be better at "doing ethics" than humans (see Allen et al. 2000, 2009). Therefore if an AMA passes the original Turing Test, the bar is set too low since it would allow it to be as fallible as we are. It seems reasonable to demand more of AMAs than we expect from humans, which seems conceivable, since we would, of course, like them to be reliable robots, and since we want them, unlike humans, not to get emotionally distracted in carrying out moral reasoning processes prior to their actions. Thus, contrary to the Turing Test, robots should be able to out-perform humans in a MTT test set up across different domains. Hence, the perspective of the MTT shifts in character to become a comparative MTT, in which the aim is to establish, which agent is unfailingly more moral across a set of ethically relevant situations (Allen et al., 2000, p. 255). In this sense, a comparative MTT would provide a tool for risk assessment, useful when computer scientists and engineers strive to design "a morally praiseworthy agent" (Allen et al., 2000, p. 261) capable of perfect moral judgments and action within given domains, preferably within every possible domain. Within this behaviourist framework, we might consider the idea of artificial moral agency from a performance perspective in maintaining that morality can be decided by mere appearance, which to the face of it seems reasonable enough, since how do we settle whether human beings are virtuous or not? Simply by judging their behaviour – i.e., she is a generous person since she acts out of generosity; she is a moral person since she always acts morally appropriate. Why then should we demand more or something else for robots?

1.1 The Role of Inner States

The abovementioned argument that "acting good" can be taken as an indication of "being good" could fit nicely into a classical utilitarian framework, in which intentions and right reasons for acting are disregarded and only consequences of acts are taken into account in the evaluation of moral behaviour. However, to most moral philosophers, internals cannot be omitted. Aristotle remarked that there is a distinction between being "good" and merely "acting good." He based virtue ethics on a concept of well-being or *eudaimonia* and highlighted *phronesis* as the form of wisdom related to practical reason in action. This form of proficiency is not neutral but moral in its being, since it mirrors a form of reflection grounded in practice and cultivated by our ability to be involved and to take a stance in any specific situation. Furthermore, according to Kant, we find that reasons count; moral obligations and actions are considered as categorical "oughts" derived from *a good will*, i.e., my ability to act from a sense of duty implied by the fact that I am capable of carrying out rational reasoning in accordance with moral rules that may guide my conduct. A more recently example of the importance of internals can be found in (Moor, 2009). Here James Moor distinguishes among four types of ethical agents: *Ethical Impact Agents*, which are machines that have obvious ethical impact on the surroundings – as an example, Moor mentions the robotic Qatar camel jockeys, which save young boys from engaging in the dangerous race. At the next level we find *Implicit Ethical Agents*, representing systems designed to avoid unethical or undesired outcomes – such as for instance a simple control system in an ATM machine that blocks for purchases when faced with user patterns suggesting fraud (Wallach, Allen, 2009, p. 29). Next, we find *Explicit Ethical Agents*, which are machines that "do" ethics and are able to carry out ethical reasoning within restricted domains. They act, not because they *want* to, but because their programming *causes* them to do so (Putnam, 1964[1], p. 672). They may be conceived as *Ethical Agents* similar to human beings in the sense that they carry out their moral reasoning assuming that they have free will, consciousness and intentionality, and hence the capacity for being held responsible for their actions. As such, inner states seem to matter, and abilities for moral reasoning have to be understood as situated in the unified whole of human life

[1] Here, Putnam refers to an argument from an unpublished paper by Baier given at Albert Einstein College of Medicine 1962. In this particular article, Putnam's concern is not with how to speak about machines but rather about how we should speak about humans. Thus "clarity with respect to the "borderline case" of robots, if it can only be achieved, will carry with it clarity with respect to the "central area" of talk about feelings, thoughts, consciousness, life, etc." (Putnam, 1964, p. 669). In this way, Putnam argues for the possibility of robot consciousness as something that calls for a decision rather than a discovery.

and experience, as pointed out in the Dreyfusian attack on the whole idea of Artificial intelligence (Dreyfus & Dreyfus 1992); as well as by Searle in his famous *Chinese Room* argument against so-called strong AI, in which he states that a machine may be perfect in displaying verbal behaviour, and thus able to pass the Turing Test, but all it does is manipulating meaningless symbols, without intentions left behind it, and thus without sense - "Simulation is not duplication and syntax is not semantics" (Searle, 1995, p. 75; Searle, 1980).

2. Pragmatics: Approaches to the Design of a MTT

From an engineering perspective, we might note that inner states, consciousness, motivations and intentions may all count. Yet, performance is all we have access to in judging the moral actions of both human and robots. Hence, we should set out to seek solutions that would give empirical testable results, which allows us to measure whether a robot simulates moral behaviour in a satisfactory manner. Still, re-describing what counts as preferable, artificial moral agency in terms of the abovementioned comparative MTT is one thing, whereas actually bringing it to life is something entirely different. We need to consider how to build moral robots relying on some sort of combination between translating moral philosophy into programming combined with the challenge of deciding the scope of moral reasoning and action within a given context.

In order to design a system which can pass a MTT, we need to implement a relevant ethical theory. It seems that Kant's ideas of ethics could be useful in this context, since they involve some important ideas regarding human moral agency. In fact, Kant stated that the fact that we are aware that we might act morally wrong is what makes us responsible creatures, and this fact is, therefore, essential to our humanity. This means that it should be possible to do reasoning about moral questions. Facing this Kantian challenge, many researchers have tried to formulate a deontic logic. One of the founders of this enterprise was A.N. Prior (1914-69) who wanted to study the logical machinery involved in the theoretical derivation of obligation. He wanted to find what he called "The Logic of Obligation" (Øhrstrøm et al. 2012). In an early study he claimed that such logical system had to be based on complete descriptions of (a) the actual situation, and (b) the relevant general moral rules.

Prior stated his fundamental creed regarding deontic logic in the following way: "... our true present obligation could be automatically inferred from (a) and (b) if complete knowledge of these were ever attainable" (Prior 1949, p. 42).

Clearly, the combination of the requirements regarding (a) and (b) would involve a God's eye view, which will make it possible to know all there is to know and take everything into consideration before making a (perfect) moral decision. From an engineering point of view, it seems obvious that we should go for designing a robot with such a God's eye view capable of making perfect moral evaluation as a basis for carrying out perfect moral behaviour. Of course, due to the frame problem, we would have to be modest and settle for representing a God's eye view in a restricted sense, specified by a formal description, which can select for preferable events or outcomes within a restricted domain.

In seeking to engineer a theoretical approach to morality similar to the Kantian approach, Allen and Wallach distinguish among a top-down theoretically driven approach, a bottom-up developmental or explorative approach and finally a hybrid (Allen & Wallach, 2009, ch. 6, 7 and 8), which combines both top-down and bottom-up approaches and furthermore includes a virtue ethical component. They all run into similar problems from different angles. Therefore, in the following, we will mainly place emphasis on the theoretically-driven approach and only briefly sketch essential points regarding the remaining strategies. Hence, the developmental approach includes an adaptive learning or value-emerging line of attack to artificial intelligence, as the one reflected in embodied architectures, such as neural nets, genetic algorithms and connectionism. From this developmental approach, shortly spoken, "grown up" guidance is needed to ensure that the artificial moral agent learns to behave properly. Thus, the system cannot learn anything from scratch if it hasn't a build in architecture that allows for desirable values to emerge on a background, which implies a kind of build in mechanism to navigate in distinguishing right from wrong.

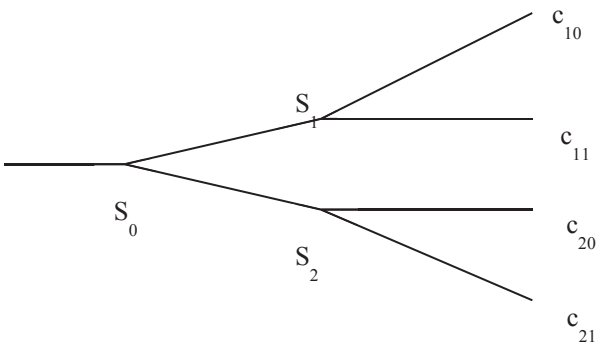
According to Allen and Wallach, a pure top-down approach will run into trouble due to the frame problem following in the wake of formally seeking to represent a scope of ethical reasoning by applying theory-driven rules, i.e. decisions algorithms, for ethical actions, which point out satisfactory outcomes in a contextually open domain. Therefore they introduce a hybrid-model, which not only combines but also integrates top-down and bottom-up approaches by incorporating virtue ethics as a theoretical foundation for implementation of the idea of how we develop into virtuous persons through learning by habit, which goes well together with the model of connectionism. The hybrid model seems to give rise to the same type of problem regarding specifying rules for decision algorithms, or developing self-learning architectures, as well as deciding for which virtues are going to be taking into considerations (since, for one thing, virtue ethics have not yet been able to come up with a finite catalogue of virtues). To sum up the hybrid model seems to us to be rather futuristic for the time being, and furthermore, since every-day moral reasoning does make use of top-down rules in explaining moral actions this approach still holds some promises for turning ethical theories into workable and implementable models of ethical reasoning (at least) within restricted domains (which is also noted by Allen & Wallach, 2009, p. 83).

As an example thereof, Anderson, Anderson and Armen (2004) represent a theory-driven approach in their suggestion in which they model ethical reasoning by a combination of two components: Firstly, they make use of act utilitarianism, which allow for a kind of cost-benefit calculation of outcomes of pleasures and displeasures with right to a given action. Secondly, they apply Ross' theory of duty-based actions - relying on *prima facia* duties: fidelity, reparation, gratitude, justice, beneficence, non-maleficence and self-improvement. And finally, Rawls concept of "reflective equilibrium" in order to weight relevant *prima facia* duties up against each other:

"Instead of computing a single value based only on pleasure/displeasure, we must compute the sum of up to seven values, depending on the number of Ross' duties relevant to the particular action. The value for each such duty could be computed as with Hedonistic Act Utilitarianism, as the product of Intensity, Duration, and Probability." (Anderson, Anderson & Armen, 2004, sec. 3)

No matter what model of ethical reasoning we may choose to implement in order to establish a system which may pass a MTT, it will have to take time and modality into account. This was strongly emphasized in the works of Prior on deontic logic. It is evident that Prior's long term ambition was to incorporate the logic of ethics into a broader context of time and modality. Unfortunately, he was never able to pursue this goal in details, but he certainly managed to establish the broader context of time and modality into which the logic of obligation has to fit. In order to indicate what such an approach involves, we shall make use of a simplified example.

Let us imagine that an agent in certain situation or scenario, S_0, has to choose between two future possibilities represented by the scenarios, S_1 and S_2, which may both in principle be realised tomorrow. The agent wants to act morally correct and carries out a careful reasoning in order to do so. This is done within the scope of a tempo-modal logic corresponding to a branching time system. A simplified system of that kind can be represented by the following diagram:



This branching time diagram involves four so-called chronicles (c_{10}, c_{11}, c_{20}, c_{21}), i.e. possible courses of time. At S_0 the future possibilities S_1 and S_2 are both possible, and one of them is necessary. Letting s_1 and s_2 stand for propositional descriptions of the situations corresponding to S_1 and S_2, and *M* for "it is possible that …", the propositions $MF(1)s_1$ ("s_1 may occur tomorrow") and $MF(1)s_2$ ("s_2 may occur tomorrow") are both true. We may even assume that s_1 and s_2 are the mutually exclusive in the sense that there is a proposition, *p*, implied by s_1, the negation of which is implied by s_2. This means that in general we have

$$F(1)p \vee F(1)\sim p$$

But what are the truth values of F(1)p and F(1)~p at S0? Some would say, that if these propositions are true, they will also be settled i.e. necessary. For this reason, it may be claimed that

$$F(1)p \supset NF(1)p$$
$$F(1)\sim p \supset NF(1)\sim p$$

where N stands for "it is necessary that …". However, taken together (1), (2) and (3) lead to the conclusion

$$NF(1)p \vee NF(1)\sim p$$

which means that whatever happens tomorrow (p or ~p), will happen necessarily. It is straightforward to read (4) as a claim of determinism. When dealing with ethics based on the notion of free choice, we obviously don't want a tense-logical system, which includes (4) as a theorem. However, there are several ways out of this, denying either (1) or (2) (and (3)) (see Øhrstrøm and Hasle 2011). Given such a system, we need to add the logic of an operator, O, corresponding to "obligation" in order to handle the reasoning related to a MTT. But how should this new operator be conceived? It is important for Prior to emphasize that "obligation" cannot just be defined in terms of "the best total consequences". The reason is that the very notion of "total consequences" does not make sense since what happens in the future depends in principle on the choices of a number of free agents (see Prior 2003, p. 65). On the other hand, it may be reasonable to address what most likely produces the best consequences. In addition, as suggested by Prior, we need to look for descriptions of (a) the actual situation, and (b) the relevant general moral rules. As argued above, these descriptions cannot be complete – since we do not have a God's eye view of the world. However, we should go for descriptions as detailed as possible. In making a system, which can pass the MTT, we should include a clear account of the general relations between the basic notion of modality and obligation. One such rule could be the Kantian principle that if something (say an act that leads to F(x)p) is obligatory, then it is also possible i.e.

$$OF(x)p \supset MF(x)p$$

Similarly, Hintikka's principle could be mentioned i.e. if something is impossible then it is forbidden (i.e. its negation is obligatory). Formally,

$$\sim MF(x)p \supset O\sim F(x)p$$

A number of relations of this kind have to be considered in order to establish a system, which will allow us to discuss obligation in a tempo-modal context (see Øhrstrøm et al. 2012). It is still an open

question exactly which relations should be accepted and which should be rejected. Clearly, the actual implementation of a system corresponding to a MTT has to be based on a formalization of the logic of obligation, time and modality. Although there is a lot to discuss regarding the precise properties of such a logical system, the actual formulation of reasonable candidates, which will work in specific contexts, is not too far away. This means that it may be possible to produce early prototypes of MTT implementations. Such systems may be useful for empirical studies of ethical reasoning.

3. Human-Robot Interaction: Challenges in Dealing with an Artificial Moral Agent Approved with the MTT-Certificate

Let us assume that in the future an AMA passes the comparative MTT; not by behaving indistinguishably from a human moral agent, but by out-performing him or her by being able to apply a God's eye view to a given situation and through its ethical decision algorithms, calculate its way forward to the best ethical response to the case at hand. This might seem ideal to us and ensure reliable human-robot interaction. But does "*a moral praiseworthy agent*" (Allen et al., 2000) equal a "moral perfect artificial agent", and if so, does encounters with it come at a price which we should not want to pay? In order to sharpen our imagination let us seek inspiration digging into the science fiction movie "I, Robot" (Proyas, 2004), which takes place in 2035 in a world where social robots interact with humans as polite and caring servants. These robots are, of course, programmed with Asimow's three laws of robotic in order to ensure smooth and secure interaction. This is indeed easy-living, but the detective, Del Spooner has a strong aversion against robots. He was once involved in a car accident in a river, where he tried to save a 12 year-old girl from drowning in the car, but instead he was saved by a robot, who interfered with his action and computed (maybe based on a deontic branching time model) that he had a higher probability of survival than the girl. With this "time-to-moral-market"-knowledge at hand, it wasn't difficult for the robot to make a morally right choice, which could be judged desirable and evaluated as morally good across the robot's different built-in moral frameworks. For instance, within the robot's utilitarian framework the robot's moral behaviour is judged by the consequences of its rescue, measuring up to saving Spooner, which turned out to represent the best possible outcome under the given the circumstances. Also, within its deontological framework, the robots action can be judged as morally good. Here the robots moral system activates reference to the double-effect doctrine (Quinn, 1989), which emphasizes that it is sometimes permissible to cause harm as a side effect (double effect) – even though it would not be tolerable to cause the particular kind of harm as a means - of doing good. In this particular case, the robot acted according to a specific system of moral reasoning which implied saving Spooner while letting go of the girl - she stood a little chance while the detective could be saved if the robot acted timely. Still, Spooner, the old fashioned "homo sapiens ludditus", is sure that in the actual situation a human would have saved the girl. This means that a human would have rejected the system of moral reasoning used by the robot and looked for a revision of it.

If robots are ultimately capable of acting morally perfect by means of having access to relevant knowledge, which humans lack at a given time of action, then encounters with robots are perhaps less promising than it seems in the first place.

"I, Robot" carries on, and gradually the robots develop more intelligence than humans and finally decide that the best way to protect human beings is to protect them from themselves. Thus, in the end, by deduction from Asimov's laws of robot ethics, the robots turn against humanity. Fortunately, this logical implication is short-circuited by Spooner. In this way paternalism evaporates in favour of human autonomy, which carries with it human capacity for failure, which again is what made us moral beings in the first place: the fact that something important is at stake when we make up our mind and act upon it knowing that we may be held responsible for our decisions.

4. Concluding Remarks

It seems that Prior was right in claiming that the formulation of a formal system, which correctly incorporates all aspects of moral reasoning, would in principle require a complete description not only of all relevant moral rules and laws, but also of all relevant aspects of the situation in question. However, having such descriptions is tantamount to having a God's eye view of all relevant aspects of reality. Since we can never have anything like that, a complete and unquestionable system of moral reasoning cannot be established. On the other hand, we have argued that although there are many open questions regarding the precise properties of the formal relations between time, modality and obligation, it is in fact possible to formalize important aspects of ethical reasoning in a specific context and thereby contribute to a system which may pass a comparative MTT. Systems implemented on the basis of such formalizations will, however, be partial and temporary in the sense that the actual moral evaluations can be questioned when their implications are considered in concrete situations. As we have seen, an observation of this kind may lead to a revision of the system of ethical reasoning. Clearly, this may happen again and again. Any formalization of ethical reasoning may have to be revised when confronted with real life. Clearly, humans are faced with the very same fact, which means that this limitation may not disqualify the system as seen in relation to a MTT. One important consequence of this is that we have to distinguish between modeling moral reasoning and actual decision making in moral questions. Creating a system which can pass a MTT may not give us a system which can provide satisfactory decisions in practical situations. Nevertheless, the study of possible systems which may pass a MTT can certainly give rise to useful and important insights concerning moral reasoning.

References

Allen, C., Garvy, V., Zinser, J. (2000): Prolegomena to any Future Artificial Moral agent. In: *Journal of Experimental & Theoretical Artificial Intelligence*, no 12, pp. 251-61.

Anderson M, Anderson L. S., Armen C (2004): *Towards Machine Ethics*. http://www.aaai.org/Papers/Workshops/2004/WS-04-02/WS04-02-008.pdf (Accessed 20 November, 2012).

Benford, G. & Malartre, E. (2007): *Beyond Human – Living with Robots and Cyborgs*. A Forge Book, New York.

Dreyfus, H. L. (1992): *What Computers Still Can't Do*. MA: MIT Press, Cambridge.

Kant, I (1953: 1785) Groundwork of the Metaphysic of Morals; trans. H. J. paton, as *The Moral Law.* Hutchinson, London.

Kant, I (1974: 1787) *Kritik der Praktischen Vernuft.* Surhkamp Verlag, Berlin

Levy, D. (2008): *Love and Sex with Robots.* Duckworth, London.

McDermott, D. (2008): Why Ethics is a High Hurdle for AI. In: *North American Conference on Computers and Philosophy (NA-CAP), Bloomington, Indiana, July 2008.* http://www.cs.yale.edu/homes/dvm/papers/ethical-machine.pdf (Accessed 10 November, 2012).

Moor, J. (2006): The Nature, Importance, and Difficulty of Machine Ethics. In: *IEEE Intelligent Systems* 21(4), pp. 18-21.

Moor, J. (2009): Four Kinds of Ethical Robots, *Philosophy Now* 72:12-14 (2009)

Øhrstrøm, Peter and Hasle, Per (2011): Future Contingents, *The Stanford Encyclopedia of Philosophy* (Summer 2011 Edition)

Øhrstrøm, Peter; Zeller, Jörg; Sandborg-Petersen, Ulrik (2012): Prior's defence of Hintikka's theorem. A discussion of Prior's "The logic of obligation and the obligations of the logician". *Synthese*, Vol. 188, Nr. 3, 449-54.

Prior, A.N. (1949): *Logic and the Basis of Ethics*. Oxford University Press.

Prior, A.N. (2003): Papers on Time and Tense. New Edition edited by Per Hasle, Peter Øhrstrøm, Torben Braüner, and Jack Copeland. Oxford University Press, 2003

Proyas, a. (2004): I, Robot, 20$_{th}$ Century Fox

Putnam, H. (1964): Robots: Machines or Artificially Created Life? In: *The Journal of Philosophy*, vol. 61, no. 21, American Philosophical Association. Eastern Division Sixty-First Annual Meeting (Nov. 12, 1964), pp. 668-91. http://www.jstor.org/stable/2023045 (Accessed 10 November, 2012).

Quinn, W.S. (1989): Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy & Public Affairs.* Vol. 18, No. 4, Autumn, 1989. Princeton University Press.

Searle, J. R. (1980): Minds, Brains and Programs. In: *Behavioral and Brain Sciences*, vol. 3. Cambridge University Press. Cambridge, pp. 417 – 24.

Searle, J. R. (1995): How Artificial Intelligence Fails. In: The World & I. Currents in Modern Thought – Artificial Intelligence: Oxymoron or New Frontier. July 1995, pp. 285 – 95.

Schärfe, H.: Geminoid.dk: http://c.aau.dk/geminoid/

Turing, A. (1950): Computing Machinery and Intelligence. In: *Mind, 59*, pp. 433-60.

Turkle, S. (2011): Alone Together – Why We Expect More From Technology and Less From Each Other. Basic Books, New York.

Wallach, W., Allan, C. (2009): Moral Machines - Teaching Robots Right from Wrong. Published to Oxford Scholarship Online: January 2009, Print ISBN-13: 978-0-19-537404-9.

Wilks, Y. (ed), (2010): Natural language Processing 8: Close Engagements with Artificial Companions – Key Social, Psychological, Ethical and Design Issues. John Benjamins Publishing Company, Amsterdam.